

Bayesian Nonparametric Modeling for Multivariate Ordinal Regression

Maria DeYoreo and Athanasios Kottas *

Abstract

Univariate or multivariate ordinal responses are often assumed to arise from a latent continuous parametric distribution, with covariate effects which enter linearly. We introduce a Bayesian nonparametric modeling approach for univariate and multivariate ordinal regression, which is based on mixture modeling for the joint distribution of latent responses and covariates. The modeling framework enables highly flexible inference for ordinal regression relationships, avoiding assumptions of linearity or additivity in the covariate effects. In standard parametric ordinal regression models, computational challenges arise from identifiability constraints and estimation of parameters requiring nonstandard inferential techniques. A key feature of the nonparametric model is that it achieves inferential flexibility, while avoiding these difficulties. In particular, we establish full support of the nonparametric mixture model under fixed cut-off points that relate through discretization the latent continuous responses with the ordinal responses. The practical utility of the modeling approach is illustrated through application to two data sets from econometrics, an example involving regression relationships for ozone concentration, and a multirater agreement problem.

KEY WORDS: Dirichlet process mixture model; Kullback-Leibler condition; Markov chain Monte Carlo; polychoric correlations.

*M. DeYoreo (maria.deyoreo@stat.duke.edu) is Postdoctoral Researcher, Department of Statistical Science, Duke University, Durham, NC, 27708, USA, and A. Kottas (thanos@ams.ucsc.edu) is Professor of Statistics, Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064, USA. This research is part of the Ph.D. dissertation of M. DeYoreo, completed at University of California, Santa Cruz, and was supported in part by the National Science Foundation under award DMS 1310438.

1 Introduction

Estimating regression relationships for univariate or multivariate ordinal responses is a key problem in many application areas. Correlated ordinal data arise frequently in the social sciences, for instance, survey respondents often assign ratings on ordinal scales (such as “agree”, “neutral”, or “disagree”) to a set of questions, and the responses given by a single rater are correlated. Ordinal data is also commonly encountered in econometrics, as rating agencies (such as Standard and Poor’s) use an ordinal rating scale. A natural way to model data of this type is to envision each ordinal variable as representing a discretized version of an underlying latent continuous random variable. In particular, the commonly used (multivariate) ordinal probit model results when a (multivariate) normal distribution is assumed for the latent variable(s).

Under the probit model for a single ordinal response Y with C categories, and covariate vector \mathbf{x} , $\Pr(Y \leq m \mid \mathbf{x}) = \Phi(\gamma_m - \mathbf{x}^T \boldsymbol{\beta})$, for $m = 1, \dots, C$. Here, $-\infty = \gamma_0 < \gamma_1 < \dots < \gamma_{C-1} < \gamma_C = \infty$ are cut-off points for the response categories, where typically $\gamma_1 = 0$ for identifiability. Albert and Chib (1993) have shown that posterior simulation is greatly simplified by augmenting the model with latent variables. In particular, assume that the ordinal response Y arises from a latent continuous response Z , such that $Y = m$ if and only if $Z \in (\gamma_{m-1}, \gamma_m]$, for $m = 1, \dots, C$, and $Z \mid \boldsymbol{\beta} \sim N(\mathbf{x}^T \boldsymbol{\beta}, 1)$, which yields $\Pr(Y = m \mid \mathbf{x}) = \int_{\gamma_{m-1}}^{\gamma_m} N(z; \mathbf{x}^T \boldsymbol{\beta}, 1) dz$.

The multivariate probit model for binary or ordinal responses generalizes the probit model to accommodate correlated ordinal responses $\mathbf{Y} = (Y_1, \dots, Y_k)$, where $Y_j \in \{1, \dots, C_j\}$ for $j = 1, \dots, k$, using a multivariate normal distribution for the underlying latent variables $\mathbf{Z} = (Z_1, \dots, Z_k)$. To obtain an identifiable model, restrictions must be imposed on the covariance matrix $\boldsymbol{\Sigma}$ of the multivariate normal distribution for \mathbf{Z} . One way to handle this is to restrict the covariance matrix to be a correlation matrix, which complicates Bayesian inference since there does not exist a conjugate prior for correlation matrices. Chib and Greenberg (1998) discuss inference under this model, using a random walk Metropolis algorithm to sample the correlation matrix. Liu (2001) and Imai and van Dyk (2005) use parameter expansion with data augmentation (Liu and Wu, 1999) to expand the parameter space such that unrestricted covariance matrices may

be sampled. Talhouk et al. (2012) work with a sparse correlation matrix arising from conditional independence assumptions, and use a parameter expansion strategy to expand the correlation matrix into a covariance matrix, which is updated with a Gibbs sampling step. Webb and Forster (2008) reparameterize Σ in a way that enables fixing its diagonal elements without sacrificing closed-form full conditional distributions. Lawrence et al. (2008) use a parameter expansion technique, in which the parameter space includes unrestricted covariance matrices, which are then normalized to correlation matrices. In addition to the challenges arising from working with correlation matrices, the setting with multivariate ordinal responses requires estimation for the cut-off points, which are typically highly correlated with the latent responses.

The assumption of normality on the latent variables is restrictive, especially for data which contains a large proportion of observations at high or low ordinal levels, and relatively few observations at moderate levels. As a consequence of the normal distribution shape, there are certain limitations on the effect that each covariate can have on the marginal probability response curves. In particular, $\Pr(Y_j = 1 \mid \mathbf{x})$ and $\Pr(Y_j = C_j \mid \mathbf{x})$ are monotonically increasing or decreasing as a function of covariate x , and they must have the opposite type of monotonicity. The direction of monotonicity changes exactly once in moving from category 1 to C_j (referred to as the single-crossing property). In addition, the relative effect of covariates r and s , i.e., the ratio of $\partial \Pr(Y_j = m \mid \mathbf{x}) / \partial x_r$ to $\partial \Pr(Y_j = m \mid \mathbf{x}) / \partial x_s$, is equal to the ratio of the r -th and s -th regression coefficients for the j -th response, which does not depend on m or \mathbf{x} . That is, the relative effect of one covariate to another is the same for every ordinal level and any covariate value. We refer to Boes and Winkelmann (2006) for further discussion of such properties.

Work on Bayesian nonparametric modeling for ordinal regression is relatively limited, particularly in the multivariate setting. In the special case of binary regression, there is only one probability response curve to be modeled, and this problem has received significant attention. Existing semiparametric approaches involve relaxing the normality assumption for the latent response (e.g., Newton et al., 1996; Basu and Mukhopadhyay, 2000), while others have targeted the linearity assumption for the response function (e.g., Mukhopadhyay and Gelfand, 1997; Walker and Mallick, 1997; Choudhuri et al., 2007). For a univariate ordinal response, Chib and

Greenberg (2010) assume that the latent response arises from scale mixtures of normals, and the covariate effects to be additive upon transformation by cubic splines. This allows nonlinearities to be obtained in the marginal regression curves, albeit under the restrictive assumption of additive covariate effects. Gill and Casella (2009) extend the ordinal probit model by introducing subject-specific random effects terms modeled with a Dirichlet process (DP) prior.

Chen and Dey (2000) model the latent variables for correlated ordinal responses with scale mixtures of normal distributions, with means linear on the covariates. In the context of multivariate ordinal data without covariates, Kottas et al. (2005) model the distribution of the latent variables with a DP mixture of multivariate normals, which is sufficiently flexible to uncover essentially any pattern in a contingency table while using fixed cut-offs. This represents a significant advantage relative to the parametric models discussed, since the estimation of cut-offs requires nonstandard inferential techniques, such as hybrid Markov chain Monte Carlo (MCMC) samplers (Johnson and Albert, 1999) and reparameterization to achieve transformed cut-offs which do not have an order restriction (Chen and Dey, 2000).

The preceding discussion reflects the fact that there are significant challenges involved in fitting multivariate probit models, and a large amount of research has been dedicated to providing new inferential techniques in this setting. While the simplicity of its model structure and interpretability of its parameters make the probit model appealing to practitioners, the assumptions of linear covariate effects, and normality on the latent variables are restrictive. Hence, from both the methodological and practical point of view, it is important to explore more flexible modeling and inference techniques. This is especially relevant for the setting with multivariate ordinal responses for which Bayesian nonparametric approaches are virtually nonexistent. Semiparametric models for binary regression are more common, since in this case there is a single regression function to be modeled. When taken to the setting involving a single ordinal response with $C \geq 3$ classifications, it becomes much harder to incorporate flexible priors for each of the $C - 1$ probability response curves. And, semiparametric prior specifications appear daunting in the multivariate ordinal regression setting where, in addition to general regression relationships, it is desirable to achieve flexible dependence structure between the ordinal responses.

Motivated by these limitations of parametric and semiparametric prior probability models, our goal is to develop a Bayesian nonparametric regression model for univariate and multivariate ordinal responses, which enables flexible inference for both the conditional response distribution and for regression relationships. We focus on problems where the covariates can be treated as random, and model the joint distribution of the covariates and latent responses with a DP mixture of multivariate normals to induce flexible regression relationships, as well as general dependence structure in the multivariate response distribution. In many fields – such as biometry, economics, and the environmental and social sciences – the assumption of random covariates is appropriate, indeed we argue necessary, as the covariates are not fixed prior to data collection.

In addition to the substantial distributional flexibility, an appealing aspect of the nonparametric modeling approach taken is that the cut-offs may be fixed, and the mixture kernel covariance matrix left unrestricted. Regarding the latter, we show that all parameters of the normal mixture kernel are identifiable provided each ordinal response comprises more than two classifications. For the former, we demonstrate that, with fixed cut-offs, our model can approximate arbitrarily well any set of probabilities on the ordinal outcomes. To this end, we prove that the induced prior on the space of mixed ordinal-continuous distributions assigns positive probability to all Kullback-Leibler (KL) neighborhoods of all densities in this space.

We are primarily interested in regression relationships, and demonstrate that we can obtain inference for a variety of nonlinear as well as more standard shapes for ordinal regression curves, using data examples chosen from fields for which the methodology is particularly relevant. As a consequence of the joint modeling framework, inferences for the conditional covariate distribution given specific ordinal response categories, or inverse inferences, are also available. These relationships may be of direct interest in certain applications. Also of interest are the associations between the ordinal variables. These are described by the correlations between the latent variables in the ordinal probit model, termed polychoric correlations in the social sciences (e.g., Olsson, 1979). Using a data set of ratings assigned to essays by multiple raters, we apply our model to determine regions of the covariate space as well as the levels of ratings at which pairs of raters tend to agree or disagree. We contrast our approach with the Bayesian methods for

studying multirater agreement in Johnson and Albert (1999) and Savitsky and Dalal (2014).

Finally, we note that our modeling framework is related to that of Shahbaba and Neal (2009), Dunson and Bhattacharya (2010), Hannah et al. (2011), and Papageorgiou et al. (2014), as they also develop nonparametric models for joint response-covariate distributions, including categorical variables. Shahbaba and Neal (2009) considered classification of a univariate response using a multinomial logit kernel, and this was extended by Hannah et al. (2011) to accommodate alternative response types with mixtures of generalized linear models. Dunson and Bhattacharya (2010) studied DP mixtures of independent kernels, and Papageorgiou et al. (2014) build a model for spatially indexed data of mixed type (count, categorical, and continuous). However, these models would not be suitable for ordinal data, or, particularly in the first three cases, when inferences are to be made on the association between ordinal response variables.

The rest of the article is organized as follows. In Section 2, we formulate the DP mixture model for ordinal regression, including study of the theoretical properties discussed above (with technical details given in the Appendix). We discuss prior specification and posterior inference, and indicate the necessary modifications when binary responses are present. The methodology is applied in Section 3 to ozone concentration data, two data examples from econometrics, and a multirater agreement problem. Finally, Section 4 concludes with a discussion.

2 Methodology

2.1 Model formulation

Suppose that k ordinal categorical variables are recorded for each of n individuals, along with p continuous covariates, so that for individual i we observe a response vector $\mathbf{y}_i = (y_{i1}, \dots, y_{ik})$ and a covariate vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, with $y_{ij} \in \{1, \dots, C_j\}$, and $C_j > 2$. Introduce latent continuous random variables $\mathbf{z}_i = (z_{i1}, \dots, z_{ik})$, $i = 1, \dots, n$, such that $y_{ij} = l$ if and only if $\gamma_{j,l-1} < z_{ij} \leq \gamma_{j,l}$, for $j = 1, \dots, k$, and $l = 1, \dots, C_j$. For example, in a biomedical application, y_{i1} and y_{i2} could represent severity of two different symptoms of patient i , recorded on a categorical scale ranging from “no problem” to “severe”, along with covariate information

weight, age, and blood pressure. The assumption that the ordinal responses represent discretized versions of latent continuous responses is natural for many settings, such as the one considered here. Note also that the assumption of random covariates is appropriate in this application, and that the medical measurements are all related and arise from some joint stochastic mechanism. This motivates our focus on building a model for the joint density $f(\mathbf{z}, \mathbf{x})$, which is a multivariate continuous density of dimension $k+p$, which in turn implies a model for the conditional response distribution $f(\mathbf{y} | \mathbf{x})$.

To model $f(\mathbf{z}, \mathbf{x})$ in a flexible way, we use a DP mixture of multivariate normals model, mixing on the mean vector and covariance matrix. That is, we assume $(\mathbf{z}_i, \mathbf{x}_i) | G \stackrel{iid}{\sim} \int N(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}) dG(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and place a DP prior (Ferguson, 1973) on the random mixing distribution G . The hierarchical model is formulated by introducing a latent mixing parameter $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for each data vector:

$$(\mathbf{z}_i, \mathbf{x}_i) | \boldsymbol{\theta}_i \stackrel{ind}{\sim} N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i); \quad \boldsymbol{\theta}_i | G \stackrel{iid}{\sim} G, \quad i = 1, \dots, n \quad (1)$$

where $G | \alpha, \boldsymbol{\psi} \sim \text{DP}(\alpha, G_0(\cdot; \boldsymbol{\psi}))$, with base distribution $G_0(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{\psi}) = N(\boldsymbol{\mu}; \mathbf{m}, \mathbf{V})\text{IW}(\boldsymbol{\Sigma}; \nu, \mathbf{S})$. The model is completed with hyperpriors on $\boldsymbol{\psi} = (\mathbf{m}, \mathbf{V}, \mathbf{S})$, and a prior on α :

$$\mathbf{m} \sim N(\mathbf{a}_m, \mathbf{B}_m), \quad \mathbf{V} \sim \text{IW}(a_V, \mathbf{B}_V), \quad \mathbf{S} \sim \text{W}(a_S, \mathbf{B}_S), \quad \alpha \sim \text{gamma}(a_\alpha, b_\alpha), \quad (2)$$

where $\text{W}(a_S, \mathbf{B}_S)$ denotes a Wishart distribution with mean $a_S \mathbf{B}_S$, and $\text{IW}(a_V, \mathbf{B}_V)$ denotes an inverse-Wishart distribution with mean $(a_V - (k+p) - 1)^{-1} \mathbf{B}_V$.

According to its constructive definition (Sethuraman, 1994), the DP generates almost surely discrete distributions, $G = \sum_{l=1}^{\infty} p_l \delta_{\boldsymbol{\theta}_l}$. The locations $\boldsymbol{\theta}_l$ are independent realizations from G_0 , and the weights are determined through stick-breaking from beta distributed random variables with parameters 1 and α . That is, $p_1 = v_1$, and $p_l = v_l \prod_{r=1}^{l-1} (1 - v_r)$, for $l = 2, 3, \dots$, with $v_l \stackrel{iid}{\sim} \text{beta}(1, \alpha)$. The discreteness of the DP allows for ties in the $\boldsymbol{\theta}_i$, so that in practice less than n distinct values for the $\{\boldsymbol{\theta}_i\}$ are imputed. The data is therefore clustered into a typically small number of groups relative to n , with the number of clusters n^* controlled by parameter α , where larger values favor more clusters (Escobar and West, 1995). From the constructive definition

for G , the prior model for $f(\mathbf{z}, \mathbf{x})$ has an almost sure representation as a countable mixture of multivariate normals, and the proposed model can therefore be viewed as a nonparametric extension of the multivariate probit model, albeit with random covariates.

This implies a countable mixture of normal distributions (with covariate-dependent weights) for $f(\mathbf{z} \mid \mathbf{x}; G)$, from which the latent \mathbf{z} may be integrated out to reveal the induced model for the ordinal regression relationships. In general, for a multivariate response $\mathbf{Y} = (Y_1, \dots, Y_k)$ with an associated covariate vector \mathbf{X} , the probability that \mathbf{Y} takes on the values $\mathbf{l} = (l_1, \dots, l_k)$, where $l_j \in \{1, \dots, C_j\}$, for $j = 1, \dots, k$, can be expressed as

$$\Pr(\mathbf{Y} = \mathbf{l} \mid \mathbf{x}; G) = \sum_{r=1}^{\infty} w_r(\mathbf{x}) \int_{\gamma_{k,l_k-1}}^{\gamma_{k,l_k}} \cdots \int_{\gamma_{1,l_1-1}}^{\gamma_{1,l_1}} N(\mathbf{z}; \mathbf{m}_r(\mathbf{x}), \mathbf{S}_r) d\mathbf{z} \quad (3)$$

with covariate-dependent weights $w_r(\mathbf{x}) \propto p_r N(\mathbf{x}; \boldsymbol{\mu}_r^x, \boldsymbol{\Sigma}_r^{xx})$, and mean vectors $\mathbf{m}_r(\mathbf{x}) = \boldsymbol{\mu}_r^z + \boldsymbol{\Sigma}_r^{zx}(\boldsymbol{\Sigma}_r^{xx})^{-1}(\mathbf{x} - \boldsymbol{\mu}_r^x)$, and covariance matrices $\mathbf{S}_r = \boldsymbol{\Sigma}_r^{zz} - \boldsymbol{\Sigma}_r^{zx}(\boldsymbol{\Sigma}_r^{xx})^{-1}\boldsymbol{\Sigma}_r^{xz}$. Here, $(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ are the atoms in the DP prior constructive definition, where $\boldsymbol{\mu}_r$ is partitioned into $\boldsymbol{\mu}_r^z$ and $\boldsymbol{\mu}_r^x$ according to random vectors \mathbf{Z} and \mathbf{X} , and $(\boldsymbol{\Sigma}_r^{zz}, \boldsymbol{\Sigma}_r^{xx}, \boldsymbol{\Sigma}_r^{zx}, \boldsymbol{\Sigma}_r^{xz})$ are the components of the corresponding partition of covariance matrix $\boldsymbol{\Sigma}_r$.

To illustrate, consider a bivariate response $\mathbf{Y} = (Y_1, Y_2)$, with covariates \mathbf{X} . The probability assigned to the event $(Y_1 = l_1) \cap (Y_2 = l_2)$ is obtained using (3), which involves evaluating bivariate normal distribution functions. However, one may be interested in the marginal relationships between individual components of \mathbf{Y} and the covariates. Referring to the example given at the start of this section, we may obtain the probability that both symptoms are severe as a function of \mathbf{X} , but also how the first varies as a function of \mathbf{X} . The marginal inference, $\Pr(Y_1 = l_1 \mid \mathbf{x}; G)$, is given by the expression

$$\sum_{r=1}^{\infty} w_r(\mathbf{x}) \left\{ \Phi \left(\frac{\gamma_{1,l_1} - m_r(\mathbf{x})}{s_r^{1/2}} \right) - \Phi \left(\frac{\gamma_{1,l_1-1} - m_r(\mathbf{x})}{s_r^{1/2}} \right) \right\} \quad (4)$$

where $m_r(\mathbf{x})$ and s_r are the conditional mean and variance for z_1 conditional on \mathbf{x} implied by the joint distribution $N(\mathbf{z}, \mathbf{x}; \boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$. Expression (4) provides also the form of the ordinal regression curves in the case of a single response.

Hence, the implied regression relationships have a mixture structure with component-specific kernels which take the form of parametric probit regressions, and weights which are covariate-dependent. This structure enables inference for non-linear, non-standard response curves, by favoring a set of parametric models with varying probabilities depending on the location in the covariate space. Many of the limitations of parametric probit models – including relative covariate effects which are constant in terms of the covariate and the ordinal level, monotonicity, and the single-crossing property of the response curves – are thereby overcome.

2.2 Model properties

In (1), Σ was left an unrestricted covariance matrix. As an alternative to working with correlation matrices under the probit model, identifiability can be achieved by fixing $\gamma_{j,2}$ (in addition to $\gamma_{j,1}$), for $j = 1, \dots, k$. As shown here, analogous results can be obtained in the random covariate setting. In particular, Lemma 1 establishes identifiability for Σ as a general covariance matrix, assuming all cut-offs, $\gamma_{j,1}, \dots, \gamma_{j,C_j-1}$, for $j = 1, \dots, k$, are fixed. Here, model identifiability refers to likelihood identifiability of the mixture kernel parameters in the induced model for (\mathbf{Y}, \mathbf{X}) , such that within a mixture component, the parameters are identifiable.

Lemma 1. *Consider a mixture of multivariate normals model, $\int N(\boldsymbol{\mu}, \Sigma) dG(\boldsymbol{\mu}, \Sigma)$, for the joint distribution of covariates \mathbf{X} and latent responses \mathbf{Z} , with fixed cut-off points defining the ordinal responses \mathbf{Y} through \mathbf{Z} . Then, the parameters $\boldsymbol{\mu}$ and Σ are identifiable in the kernel of the implied mixture model for (\mathbf{Y}, \mathbf{X}) , provided $C_j > 2$, for $j = 1, \dots, k$.*

Refer to Appendix A for a proof of this result. If $C_j = 2$ for some j , additional restrictions are needed for identifiability, as discussed in Section 2.5.

Identifiability is a basic model property, and is achieved here by fixing the cut-offs. However, this may appear to be a significant restriction, as under the parametric probit setting, fixing all cut-off points would prohibit the model from being able to adequately assign probabilities to the regions determined by the cut-offs. We therefore seek to determine if the nonparametric model with fixed cut-offs is sufficiently flexible to accommodate any distribution for (\mathbf{Y}, \mathbf{X}) and also for $\mathbf{Y} \mid \mathbf{X}$. Kottas et al. (2005) provide an informal argument that the normal DP mixture

model for multivariate ordinal responses (without covariates) can approximate arbitrarily well any probability distribution for a contingency table. The basis for this argument is that, in the limit, one mixture component can be placed within each set of cut-offs corresponding to a specific ordinal vector, with the mixture weights assigned accordingly to each cell.

Here, we provide a proof of the full support of our more general model for ordinal-continuous distributions. In addition to being a desirable property on its own, the ramifications of full support for the prior model are significant, as it is a key condition for the study of posterior consistency (e.g., Ghosh and Ramamoorthi, 2003). Using the KL divergence to define density neighborhoods, a particular density $f_0(\mathbf{w})$ is in the KL support of the prior \mathcal{P} , if $\mathcal{P}(K_\epsilon(f_0(\mathbf{w}))) > 0$, for any $\epsilon > 0$, where $K_\epsilon(f_0(\mathbf{w})) = \{f : \int f_0(\mathbf{w}) \log(f_0(\mathbf{w})/f(\mathbf{w}))d\mathbf{w} < \epsilon\}$. The KL property is satisfied if any density in the space of interest is in the KL support of the prior.

It has been established that the DP location normal mixture prior model satisfies the KL property (Wu and Ghosal, 2008). That is, if the mixing distribution G is assigned a DP prior on the space of random distributions for $\boldsymbol{\mu}$, and a normal kernel is chosen such that $f(\mathbf{w}; G, \boldsymbol{\Sigma}) = \int N(\mathbf{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})dG(\boldsymbol{\mu})$, with $\boldsymbol{\Sigma}$ a diagonal matrix, then the induced prior model for densities has full KL support. Letting this induced prior be denoted by \mathcal{P} , and modeling the joint distribution of (\mathbf{X}, \mathbf{Z}) with a DP location mixture of normals, the KL property yields:

$$\mathcal{P}\left(\left\{f : \int f_0(\mathbf{x}, \mathbf{z}) \log(f_0(\mathbf{x}, \mathbf{z})/f(\mathbf{x}, \mathbf{z}))d\mathbf{x}d\mathbf{z} < \epsilon\right\}\right) > 0 \quad (5)$$

for any $\epsilon > 0$, and all densities $f_0(\mathbf{x}, \mathbf{z}) \in \mathcal{D}$, where \mathcal{D} denotes the space of densities on \mathbb{R}^{p+k} .

To establish the KL property of the prior on mixed ordinal-continuous distributions for (\mathbf{X}, \mathbf{Y}) induced from multivariate continuous distributions for (\mathbf{X}, \mathbf{Z}) , consider a generic data-generating distribution $p_0(\mathbf{x}, \mathbf{y}) \in \mathcal{D}^*$. Here, \mathcal{D}^* denotes the space of distributions on $\mathbb{R}^p \times \{1, \dots, C_1\} \times \dots \times \{1, \dots, C_k\}$. Let $f_0(\mathbf{x}, \mathbf{z})$ be a density function on \mathbb{R}^{p+k} such that

$$p_0(\mathbf{x}, l_1, \dots, l_k) = \int_{\gamma_{k, l_k-1}}^{\gamma_{k, l_k}} \dots \int_{\gamma_{1, l_1-1}}^{\gamma_{1, l_1}} f_0(\mathbf{x}, z_1, \dots, z_k) dz_1 \dots dz_k \quad (6)$$

for any (l_1, \dots, l_k) , where $l_j \in \{1, \dots, C_j\}$, for $j = 1, \dots, k$. That is, $f_0(\mathbf{x}, \mathbf{z})$ is a latent continuous

density which induces the corresponding distribution on the ordinal responses. Note that at least one $f_0 \in \mathcal{D}$ exists for each $p_0 \in \mathcal{D}^*$, with one such f_0 described in Appendix B. The next lemma (proved in Appendix B) establishes that, as a consequence of the KL property of the DP mixture of normals (5), the prior assigns positive probability to all KL neighborhoods of $p_0(\mathbf{x}, \mathbf{y})$, as well as to all KL neighborhoods of the implied conditional distributions $p_0(\mathbf{y} \mid \mathbf{x})$.

Lemma 2. *Assume the distribution of a mixed ordinal-continuous random variable is $p_0(\mathbf{x}, \mathbf{y}) \in \mathcal{D}^*$, and let $f_0(\mathbf{x}, \mathbf{z}) \in \mathcal{D}$ be the corresponding continuous density function, for which $\mathcal{P}(K_\epsilon(f_0(\mathbf{x}, \mathbf{z}))) > 0$, for any $\epsilon > 0$. Then $\mathcal{P}(K_\epsilon(p_0(\mathbf{x}, \mathbf{y}))) > 0$ and $\mathcal{P}(K_\epsilon(p_0(\mathbf{y} \mid \mathbf{x}))) > 0$, for any $\epsilon > 0$.*

Lemma 2 establishes full KL support for a model arising from a DP location normal mixture, a simpler version of our model. Combined together, the properties of identifiability and full support reflect a major advantage of the proposed model. That is, it can approximate arbitrarily well any distribution on (\mathbf{Y}, \mathbf{X}) , as well as any conditional distribution for $(\mathbf{Y} \mid \mathbf{X})$, while at the same time avoiding the need to impute cut-off points or work with correlation matrices, both of which are major challenges in fitting multivariate probit models.

The cut-offs can be fixed to arbitrary increasing values, which we recommend to be equally spaced and centered at zero. As confirmed empirically with simulated data (refer to Section 3 for details), the choice of cut-offs does not affect inferences for the ordinal regression relationships, only the center and scale of the latent variables, which must be interpreted relative to the cut-offs.

2.3 Prior specification

To implement the model, we need to specify the parameters of the hyperpriors in (2). A default specification strategy is developed by considering the limiting case of the model as $\alpha \rightarrow 0^+$, which results in a single normal distribution for (\mathbf{Z}, \mathbf{X}) . This limiting model is essentially the multivariate probit model, with the addition of random covariates. The only covariate information we use here is an approximate center and range for each covariate, denoted by $\mathbf{c}^x = (c_1^x, \dots, c_p^x)$ and $\mathbf{r}^x = (r_1^x, \dots, r_p^x)$. Then c_m^x and $r_m^x/4$ are used as proxies for the marginal mean and standard deviation of X_m . We also seek to center and scale the latent variables appropriately, using the cut-offs. Since Y_j is supported on $\{1, \dots, C_j\}$, latent continuous variable Z_j must be

supported on values slightly below $\gamma_{j,1}$, up to slightly above γ_{j,C_j-1} . We therefore use $r_j^z/4$, where $r_j^z = (\gamma_{j,C_j-1} - \gamma_{j,1})$, as a proxy for the standard deviation of Z_j .

Under $(\mathbf{Z}, \mathbf{X}) \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have $\mathbb{E}(\mathbf{Z}, \mathbf{X}) = \mathbf{a}_m$, and $\text{Cov}(\mathbf{Z}, \mathbf{X}) = a_S \mathbf{B}_S(\nu - d - 1)^{-1} + \mathbf{B}_V(a_V - d - 1)^{-1} + \mathbf{B}_m$, with $d = p + k$. Then, assuming each set of cut-offs $(\gamma_{j,0}, \dots, \gamma_{j,C_j})$ are centered at 0, we fix $a_m = (0, \dots, 0, \mathbf{c}^x)$. Letting $\mathbf{D} = \text{diag}\{(r_1^z/4)^2, \dots, (r_k^z/4)^2, (r_1^x/4)^2, \dots, (r_p^x/4)^2\}$, each of the three terms in $\text{Cov}(\mathbf{Z}, \mathbf{X})$ can be assigned an equal proportion of the total covariance, and set to $(1/3)\mathbf{D}$, or to $(1/2)\mathbf{D}$ to inflate the variance slightly. For dispersed but proper priors with finite expectation, ν , a_V , and a_S can be fixed to $d + 2$. Fixing these parameters allows for \mathbf{B}_S and \mathbf{B}_V to be determined accordingly, completing the default specification strategy for the hyperpriors of \mathbf{m} , \mathbf{V} , and \mathbf{S} .

In the strategy outlined above, the form of $\text{Cov}(\mathbf{Z}, \mathbf{X})$ was diagonal, such that a priori we favor independence between \mathbf{Z} and \mathbf{X} within a particular mixture component. Combined with the other aspects of the prior specification approach, this generally leads to prior means for the ordinal regression curves which do not have any trend across the covariate space. Moreover, the corresponding prior uncertainty bands span a significant portion of the unit interval.

2.4 Posterior inference

The approach to MCMC posterior simulation is based on a finite truncation approximation to the countable mixing distribution G , using the DP stick-breaking representation. The blocked Gibbs sampler (Ishwaran and Zarepour, 2000; Ishwaran and James, 2001) replaces the countable sum with a finite sum, $G_N = \sum_{l=1}^N p_l \delta_{(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$, with $(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ i.i.d. from G_0 , for $l = 1, \dots, N$. Here, the first $N - 1$ elements of $\mathbf{p} = (p_1, \dots, p_N)$ are equivalent to those in the countable representation of G , whereas $p_N = 1 - \sum_{l=1}^{N-1} p_l$. Under this approach, the posterior samples for model parameters yield posterior samples for G_N , and therefore full inference is available for mixture functionals.

To express the hierarchical model for the data after replacing G with G_N , introduce configuration variables (L_1, \dots, L_n) , such that $L_i = l$ if and only if $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$, for $i = 1, \dots, n$ and

$l = 1, \dots, N$. Then, the model for the data becomes

$$\begin{aligned}
y_{ij} = l \quad &\text{iff} \quad \gamma_{j,l-1} < z_{ij} \leq \gamma_{j,l}, \quad i = 1, \dots, n, \quad j = 1, \dots, k \\
(\mathbf{z}_i, \mathbf{x}_i) \mid &\{(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) : l = 1, \dots, N\}, L_i \stackrel{\text{ind.}}{\sim} N(\boldsymbol{\mu}_{L_i}, \boldsymbol{\Sigma}_{L_i}), \quad i = 1, \dots, n \\
L_i \mid \mathbf{p} &\stackrel{\text{iid}}{\sim} \sum_{l=1}^N p_l \delta_l(L_i), \quad i = 1, \dots, n \\
(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \mid &\boldsymbol{\psi} \stackrel{\text{iid}}{\sim} N(\boldsymbol{\mu}_l; \mathbf{m}, \mathbf{V}) \text{IW}(\boldsymbol{\Sigma}_l; \nu, \mathbf{S}), \quad l = 1, \dots, N
\end{aligned}$$

where the prior density for \mathbf{p} is given by $\alpha^{N-1} p_N^{\alpha-1} (1 - p_1)^{-1} (1 - (p_1 + p_2))^{-1} \times \dots \times (1 - \sum_{l=1}^{N-2} p_l)^{-1}$, which is a special case of the generalized Dirichlet distribution. The full model is completed with the conditionally conjugate priors on $\boldsymbol{\psi}$ and α as given in (2).

All full posterior conditional distributions are readily sampled, enabling efficient Gibbs sampling from the joint posterior distribution of the model above. Conditional on the latent responses \mathbf{z}_i , we have standard updates for the parameters of a normal DP mixture model. And conditional on the mixture model parameters, each z_{ij} , for $i = 1, \dots, n$ and $j = 1, \dots, k$, has a truncated normal full conditional distribution supported on the interval $(\gamma_{j,y_{ij}-1}, \gamma_{j,y_{ij}}]$.

The regression functional $\Pr(\mathbf{Y} = \mathbf{l} \mid \mathbf{x}; G)$ (estimated by the truncated version of (3) implied by G_N) can be computed over a grid in \mathbf{x} at every MCMC iteration. This yields an entire set of samples for ordinal response probabilities at any covariate value \mathbf{x} (note that \mathbf{x} may include just a portion of the covariate vector or a single covariate). As indicated in (4), in the multivariate setting, we may wish to report inference for individual components of \mathbf{Y} over the covariate space.

In some applications, in addition to modeling how \mathbf{Y} varies across \mathbf{X} , we may also be interested in how the distribution of \mathbf{X} changes at different ordinal values of \mathbf{Y} . As a feature of the modeling approach, we can obtain inference for $f(\mathbf{x} \mid \mathbf{y}; G)$, for any configuration of ordinal response levels \mathbf{y} . We refer to these inferences as inverse relationships, and illustrate them with the data example of Section 3.2.

Under the multivariate response setting, the association between the ordinal variables may also be a key target of inference. In the social sciences, the correlations between pairs of latent responses, $\text{corr}(Z_r, Z_s)$, are termed polychoric correlations (Olsson, 1979) when a single

multivariate normal distribution is used for the latent responses. Under our mixture modeling framework, we can sample a single $\text{corr}(Z_r, Z_s)$ at each MCMC iteration according to the corresponding \mathbf{p} , providing posterior predictive inference to assess overall agreement between pairs of response variables. As an alternative, and arguably more informative measure of association, we can obtain inference for probability of agreement over each covariate, or probability of agreement at each ordinal level. These inferences can be used to identify parts of the covariate space where there is agreement between response variables, as well as the ordinal values which are associated with higher levels of agreement. In the social sciences it is common to assess agreement among multiple raters or judges who are assigning a grade to the same item. In Section 3.4, we illustrate our methods with such a multirater agreement data example, in which both estimating regression relationships and modeling agreement are major objectives.

2.5 Accommodating binary responses

Our methodology focuses on multivariate ordinal responses with $C_j > 2$, for all j . However, if one or more responses is binary, then the full covariance matrix of the normal mixture kernel for (\mathbf{Z}, \mathbf{X}) is not identifiable. In the univariate probit model, it is standard practice to assume that $Z \mid \beta \sim N(\mathbf{x}^T \beta, 1)$, that is, identifiability is achieved by fixing Σ^{zz} . DeYoreo and Kottas (2014) show that a similar restriction suffices in binary regression inference built through normal mixture modeling for the joint distribution of the covariates and the single latent response.

Under the general setting involving some binary and some ordinal responses, extending the argument that led to Lemma 1, it can be shown that identifiability is accomplished by fixing the diagonal elements of Σ^{zz} that correspond to the variances of the latent binary responses (the associated covariance elements remain identifiable). To incorporate this restriction, the inverse Wishart distribution for the component of G_0 that corresponds to Σ must be replaced with a more structured specification. To this end, a square-root-free Cholesky decomposition of Σ (Daniels and Pourahmadi, 2002; Webb and Forster, 2008) can be used to fix Σ^{zz} in the setting with a single binary response (DeYoreo and Kottas, 2014), and this is also useful in the multivariate setting. This decomposition expresses Σ in terms of a unit lower triangular

matrix \mathbf{B} , and a diagonal matrix $\mathbf{\Delta}$, with elements $\delta_i > 0$, $i = 1, \dots, k + p$, such that $\mathbf{\Sigma} = \mathbf{B}^{-1}\mathbf{\Delta}(\mathbf{B}^{-1})^T$. The key result is that if $(W_1, \dots, W_r) \sim N(\boldsymbol{\mu}, \mathbf{B}^{-1}\mathbf{\Delta}(\mathbf{B}^{-1})^T)$, then this implies $\text{Var}(W_i \mid W_1, \dots, W_{i-1}) = \delta_i$, for $i = 2, \dots, r$. Therefore, if $(\mathbf{Z}, \mathbf{X}) \sim N(\boldsymbol{\mu}, \mathbf{B}^{-1}\mathbf{\Delta}(\mathbf{B}^{-1})^T)$, with (Z_1, \dots, Z_r) binary, and (Z_{r+1}, \dots, Z_k) ordinal, then fixing δ_1 fixes $\text{Var}(Z_1)$, fixing δ_2 fixes $\text{Var}(Z_2 \mid Z_1)$, and so on. The scale of the latent binary responses may therefore be constrained by fixing δ_1 , the variance of the first latent binary response, Z_1 , and the conditional variances $(\delta_2, \dots, \delta_r)$ of the remaining latent binary responses (Z_2, \dots, Z_r) . The conditional variances $(\delta_{r+1}, \dots, \delta_{k+p})$ are not restricted, since they correspond to the scale of latent ordinal responses or covariates, which are identifiable under our model with fixed cut-offs.

3 Data Examples

The model was extensively tested on simulated data, where the primary goal was to assess how well it can estimate regression functionals which exhibit highly non-linear trends. We also explored effects of sample size, choice of cut-offs, and number of response categories.

The effect of sample size was observed in the uncertainty bands for the regression functions, which were reduced in width and became smoother with a larger sample size. The regression estimates captured the truth well in all cases, but were smoother and more accurate with more data, as expected. Even with a five-category ordinal response and only 200 observations, the model was able to capture quite well non-standard regression curves.

As discussed previously, the cut-offs may be fixed to arbitrary increasing values, with the choice expected to have no impact on inference involving the relationship between \mathbf{Y} and \mathbf{X} . To test this point, the model was fit to a synthetic data set containing an ordinal response with three categories, using cut-off points of $(-\infty, -20, 20, \infty)$ and $(-\infty, -5, 5, \infty)$, as well as cut-offs $(-\infty, 0, 0.1, \infty)$, which correspond to a narrow range of latent Z values producing response value $Y = 2$. The ordinal regression function estimates were unaffected by the change in cut-offs. The last set of cut-offs forces the model to generate components with small variance (lying in the interval $(0, 0.1)$), but the resulting regression estimates are unchanged from the previous ones.

In the data illustrations that follow, the default prior specification strategy outlined in Section 2.3 was used. The posterior distributions for each component of \mathbf{m} were always very peaked compared to the prior. Some sensitivity to the priors was found in terms of posterior learning for hyperparameters \mathbf{V} and \mathbf{S} , however this was not reflected in the posterior inferences for the regression functions, which displayed little to no change when the priors were altered. Regarding the DP precision parameter α , we noticed a moderate amount of learning for larger data sets, and a small amount for smaller data sets, which is consistent with what has been empirically observed about α in DP mixtures. The prior for α was in all cases chosen to favor reasonably large values, relative to the sample size, for the number of distinct mixture components.

3.1 Ozone data

Problems in the environmental sciences provide a broad area of application for which the proposed modeling approach is particularly well-suited. For such problems, it is of interest to estimate relationships between different environmental variables, some of which are typically recorded on an ordinal scale even though they are, in fact, continuous variables. This is also a setting where it is natural to model the joint stochastic mechanism for all variables under study from which different types of conditional relationships can be explored.

To illustrate the utility of our methodology in this context, we work with data set `ozone` from the “ElemStatLearn” R package. This example contains four variables: ozone concentration (ppb), wind speed (mph), radiation (langleys), and temperature (degrees Fahrenheit). While these environmental characteristics are all random and interrelated, we focus on estimating ozone concentration as a function of radiation, wind speed, and temperature. To apply our model, rather than using directly the observed ozone concentration, we define an ordinal response containing three ozone concentration categories. Ozone concentration greater than 100 ppb is defined as high (ordinal level 3); this can be considered an extreme level of ozone concentration, as only about 6% of the total of 111 observations are this high. Concentration falling between 50 ppb and 100 ppb (approximately 25% of the observations) is considered medium (ordinal level 2), and values less than 50 ppb are assumed low (ordinal level 1).

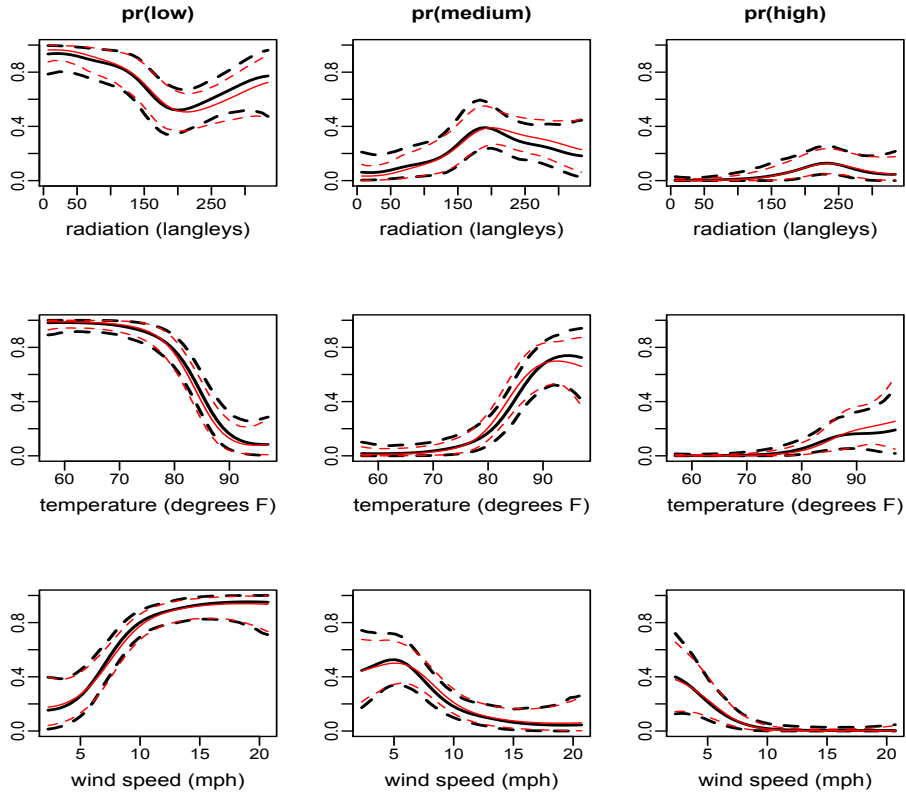


Figure 1: Ozone data. Posterior mean (solid lines) and 95% interval estimates (dashed lines) for $\Pr(Y = l \mid x_m; G)$ (thick black) compared to $\Pr(\gamma_{l-1} < Z \leq \gamma_l \mid x_m; G)$ (red), for $l = 1, 2, 3$ and $m = 1, 2, 3$, giving the probability that ozone concentration is low, medium, and high over covariates radiation, temperature, and wind speed.

The model was applied to the ozone data, with response, Y , given by discretized ozone concentration, and covariates, $\mathbf{X} = (X_1, X_2, X_3) = (\text{radiation}, \text{temperature}, \text{wind speed})$. To validate the inferences obtained from the DP mixture ordinal regression model, we compare results with the ones from a DP mixture of multivariate normals for the continuous vector (Z, \mathbf{X}) , since the observations for ozone concentration, Z , are available on a continuous scale. The latter model corresponds to the curve-fitting approach to regression of Müller et al. (1996), extended with respect to the resulting inferences in Taddy and Kottas (2010). Here, it serves as a benchmark for our ordinal regression model, since it provides the best possible inference that can be obtained under the mixture modeling framework if no loss in information occurs by observing Y rather than Z . We compare the univariate regression curves $\Pr(Y = l \mid x_m; G)$ with

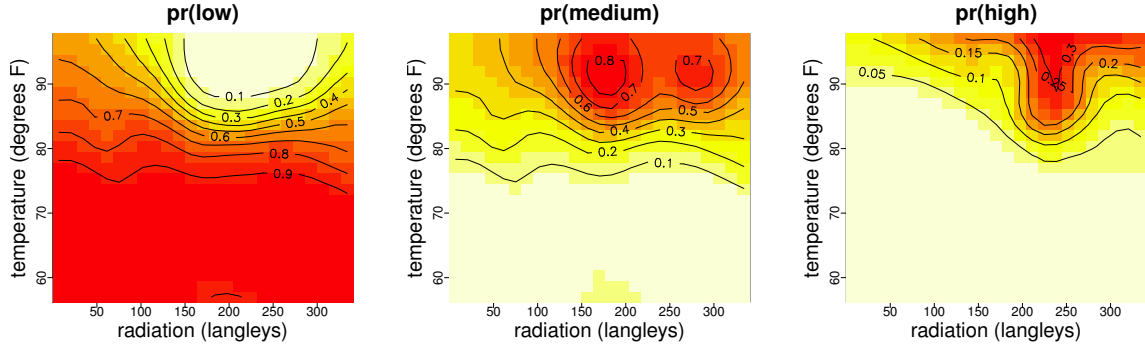


Figure 2: Ozone data. Posterior mean estimates for $\Pr(Y = l \mid x_1, x_2; G)$, for $l = 1, 2, 3$, corresponding to, from left to right, low, medium, and high ozone concentration. The spectrum of colors from white to red indicates probabilities that range from 0 to 1.

$\Pr(\gamma_{l-1} < Z \leq \gamma_l \mid x_m; G)$, for $l = 1, 2, 3$, and $m = 1, 2, 3$, the latter based on the mixture model for (Z, \mathbf{X}) . Figure 1 compares posterior mean and 95% interval estimates for the regression curves given each of the three covariates. The key result is that both sets of inferences uncover the same regression relationship trends. The only subtle differences are in the uncertainty bands which are overall slightly wider under the ordinal regression model. Also noteworthy is the fact that the ordinal regression mixture model estimates both relatively standard monotonic regression functions (e.g., for temperature) as well as non-linear effects for radiation.

The ability to capture such a wide range of trends for the regression relationships is a feature of the nonparametric mixture model. Another feature is its capacity to accommodate interaction effects among the covariates without the need to incorporate additional terms in the model. Such effects are suggested by the estimates for the response probability surfaces over pairs of covariates; for instance, Figure 2 displays these estimates as a function of radiation and temperature.

3.2 Credit ratings of U.S. companies

Here, we consider an example involving Standard and Poor's (S&P) credit ratings for 921 U.S. firms in 2005. The example is taken from Verbeek (2008), in which an ordered logit model was applied to the data, and was also used by Chib and Greenberg (2010) to illustrate an additive cubic spline regression model with a normal DP mixture error distribution. For each firm, a

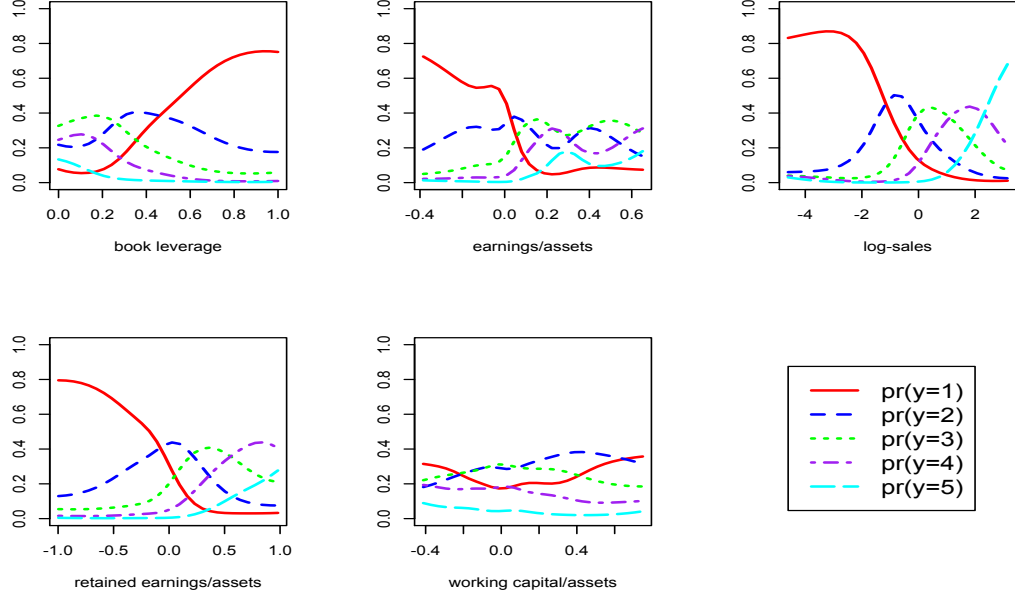


Figure 3: Credit rating data. Posterior mean estimates for $\Pr(Y = l \mid x_m; G)$, for each covariate $m = 1, \dots, 5$. All five ordinal response curves are displayed in a single panel for each covariate.

credit rating on a seven-point ordinal scale is available, along with five characteristics. Consistent with the analysis of Chib and Greenberg (2010), we combined the first two categories as well as the last two categories to produce an ordinal response with 5 levels, where higher ratings indicate more creditworthiness. The covariates in this application are book leverage X_1 (ratio of debt to assets), earnings before interest and taxes divided by total assets X_2 , standardized log-sales X_3 (proxy for firm size), retained earnings divided by total assets X_4 (proxy for historical profitability), and working capital divided by total assets X_5 (proxy for short-term liquidity).

The posterior mean estimates for the marginal probability curves, $\Pr(Y = l \mid x_m; G)$, for $l = 1, \dots, 5$ and $m = 1, \dots, 5$, are shown in Figure 3. These estimates depict some differences from the corresponding ones reported in Chib and Greenberg (2010), which could be due to the additivity assumption of the covariate effects in the regression function under their model. Empirical regression functions – computed by calculating proportions of observations assigned to each class over a grid in each covariate – give convincing graphical evidence that the regression relationships estimated by our model fit the data quite well.

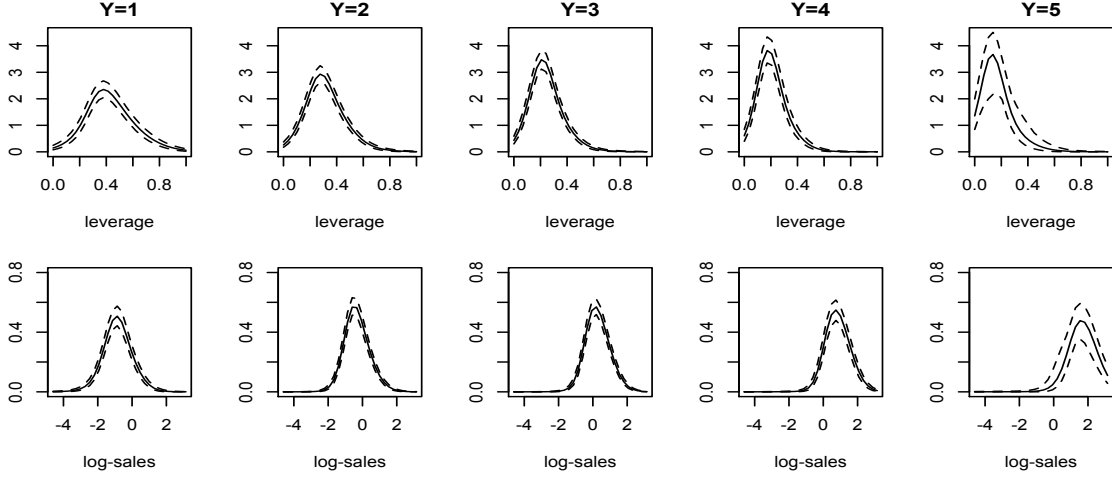


Figure 4: Credit rating data. Posterior mean (solid lines) and 95% interval estimates (dashed lines) for covariate densities $f(x | Y = l; G)$ conditional on ordinal credit rating $l = 1, \dots, 5$. The top row corresponds to covariate book leverage, and the bottom row to standardized log-sales.

To discuss the estimated regression trends for one of the covariates, consider the standardized log-sales variable, which is a proxy for firm size. The probability of rating level 1 is roughly constant for low log-sales values, and is then decreasing to 0, indicating that small firms have a similar high probability of receiving the lowest rating, whereas the larger the firm, the closer to 0 this probability becomes. The probability curves for levels 2, 3, and 4 are all quadratic shaped, with peaks occurring at larger log-sales values for higher ratings. Finally, the probability of receiving the highest rating is very small for low to moderate log-sales values, and is increasing for larger values. In summary, log-sales are positively related to credit rating, as expected.

This is another example where it is arguably natural to model the joint distribution for the response and covariates (the specific firm characteristics). This allows our model to accommodate interactions between covariates, as we do not assume additivity or independence in the effects of the covariates on the latent response. In addition to the regression curve estimates, we may obtain inference for the covariate distribution, or for any covariate conditional on a specific ordinal rating. These inverse relationships (discussed in Section 2.4) could be practically relevant in this application. It may be of interest to investors and econometricians to know, for example, approximately how large is a company's leverage, given that it has a rating of 2? Is

the distribution of leverage much different from that of a level 3 company? Figure 4 plots the estimated densities for book leverage and standardized log-sales conditional on each of the five ordinal ratings. In general, the distribution of book leverage is centered on smaller values as rating increases, and the densities become more peaked supporting a smaller range of leverage values for higher ratings. The interval bands are slightly wider for the distribution associated with $Y = 1$ than for $Y = 2, 3$, or 4 , and they are much wider for $Y = 5$, which is consistent with the small number of firms with a rating of 5. The distribution of log-sales has a mode which occurs at increasing values as rating increases, indicating that if one firm has a higher rating than another, it likely also has higher sales.

3.3 Standard and Poor's grades of countries

As a second econometrics example, we consider a data set from Simonoff (2003), comprising S&P ratings of $n = 31$ countries along with debt service ratio and income, the latter recorded on an ordinal scale with levels of low, medium, and high. Ratings range from 1 to 7, with 1 indicating the best rating of AAA, and 7 the worst of CCC. With two covariates and a very small sample size, this example provides an interesting testbed for our modeling approach.

Since income is available as a discrete variable, W , we model it through the (latent) continuous income variable, Z_2 . Therefore, X represents debt service ratio, and $\mathbf{Z} = (Z_1, Z_2)$, where W arises from Z_2 just as the ordinal rating response Y arises through latent continuous response Z_1 . This is another application where one may be interested in inverse relationships, such as the distribution of debt service ratio and/or income given a specific S&P rating.

The probability response curves as functions of debt service ratio (Figure 5) contain both monotonic trends (decreasing for response categories 1 and 2, and increasing for category 7), as well as non-linear ones, most notably for categories 4 and 5. The interval bands are wider than in earlier examples, given the smaller sample size. Although not shown here, regression curves can also be obtained over discrete income from $\Pr(Y = j \mid W = w; G) = \Pr(Y = j, W = w; G) / \Pr(W = w; G)$, for $w = 1, 2, 3$ (low, medium, and high income), where the numerator contains a double integral of a bivariate normal density function. The probability of receiving a

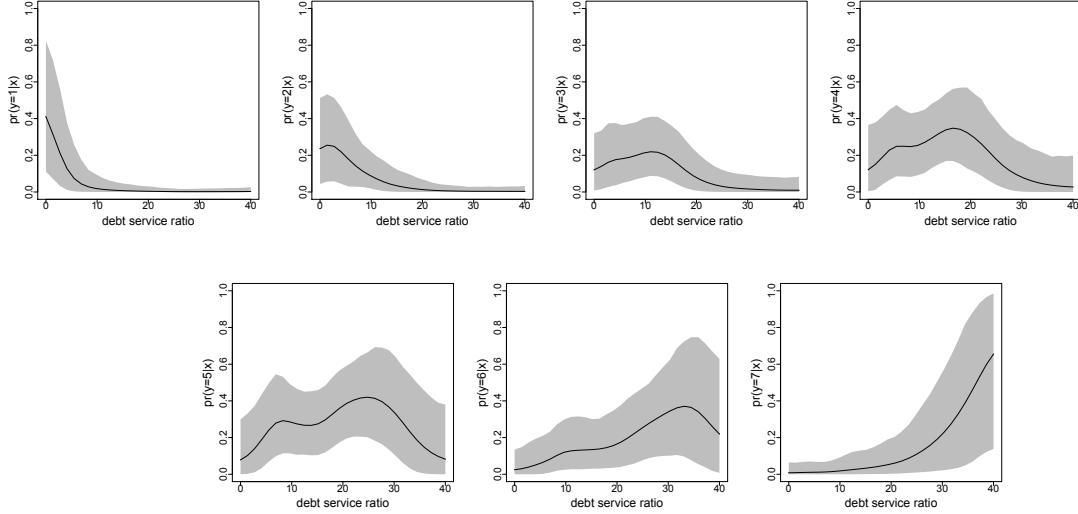


Figure 5: S&P ratings of countries data. Posterior mean (solid lines) and 95% intervals (gray bands) for the probability response curve associated with each rating, ranging from AAA (level 1) to CCC (level 7), as a function of debt service ratio.

top rating of 1, 2, or 3 is highest for high-income countries, the probability of receiving a moderate rating of 4 or 5 is highest for medium-income countries, and the probability of receiving a poor rating is highest for low-income countries. It is highly unlikely for a country to receive one of the top two ratings unless it is high-income, however there is non-negligible probability of a medium-income country receiving one of the two lowest ratings.

The latent continuous responses represent latent continuous credit rating in this application. The method for posterior simulation involves sampling $z_{i,1}$, for $i = 1, \dots, 31$, which represent the country-specific latent ratings. The two countries with AA (level 2) rating are Canada and Australia. Both of these countries have income classified as high, however Canada has no debt, whereas Australia has a debt service ratio which is around 10. This value is not particularly high, but since higher debt service ratio seems to be associated with poorer ratings, we would expect that Canada would be closer to receiving a better rating of AAA than Australia. Indeed, posterior densities of latent continuous ratings for Canada and Australia (Figure 6, left panel) indicate that Canada's ordinal rating of AA is closer to a AAA rating than Australia's.

Next, consider the four countries with A rating (level 3): Chile, Czech Republic, Hungary,

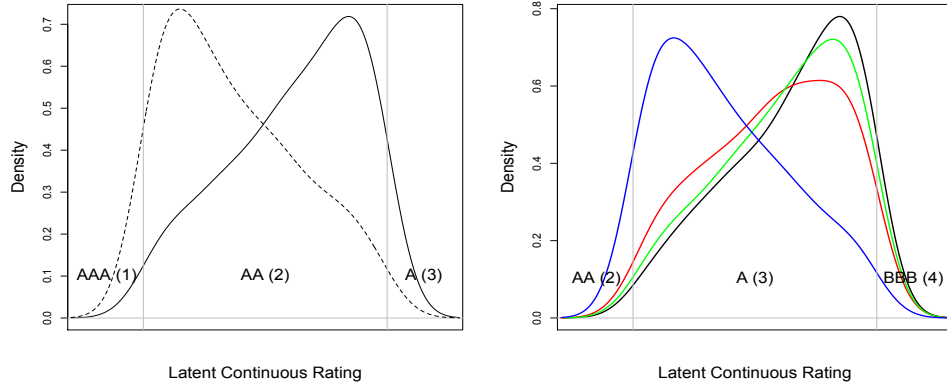


Figure 6: S&P ratings of countries data. Left panel: Posterior densities of latent continuous ratings for Australia (solid line) and Canada (dashed line), the two countries with AA rating. Right panel: Posterior densities of latent continuous ratings for the four countries with A rating: Chile (black), Czech Republic (red), Hungary (green), and Slovenia (blue). In both panels, the gray vertical lines indicate the borders for ordinal ratings.

and Slovenia. The first three of these countries are classified as medium income, whereas Slovenia is a high income country. The debt service ratios range from 8.9 (Czech Republic) to 15.8 (Chile). The estimated latent response densities are shown on the right panel of Figure 6. We note that Slovenia's latent rating distribution is centered on values close to the cut-off point for a better rating of AA. The other three distributions are fairly similar. Chile appears closest to a BBB rating, which is consistent with its higher debt service ratio. An interesting observation from these results is that differences in income appear to have a larger effect on the latent rating distributions than differences in debt service ratio.

3.4 Analysis of multirater agreement data

A variety of methods exist for analyzing ordinal data collected from multiple raters when the goal is to measure agreement. Such methods range from the commonly used κ statistic (Cohen, 1960) and its extensions (Fleiss, 1971), which are indices calculated from the observed and expected agreement of raters, to model-based approaches involving log-linear models (Tanner and Young, 1985). We do not attempt a comprehensive review here, rather our focus lies in the use of model-based methods for ordinal responses collected from multiple raters along with covariate

information. The proposed multivariate ordinal regression model offers flexibility in this setting in terms of the modeling framework and resulting inferences. We focus on a scenario involving a set of expert graders who evaluate student essays, rating them on an ordinal scale. We contrast our approach to the parametric model of Johnson and Albert (1999), from where the specific data example is taken, and the semiparametric approach of Savitsky and Dalal (2014), both developing Bayesian inference built from modeling for latent responses.

Multirater agreement data arises when k raters assign ordinal scores to n individuals, such that $\mathbf{y}_i = (y_{i1}, \dots, y_{ik})$ collects all scores for the i th individual. The raters typically use the same classification levels, and therefore each $y_{ij} \in \{1, \dots, C\}$. This data structure could be summarized in a contingency table, however, we are concerned with problems in which relevant covariate information is available for each individual. We assume that each judge assigns an ordinal rating to individual i , which represents a discretized version of a continuous rating; that is, y_{ij} is determined by z_{ij} , the continuous latent score assigned by judge j on individual i .

This is in contrast to the formulation of Johnson and Albert (1999), where all judges are assumed to agree on the intrinsic worth of each item, such that $z_{ij} = w_i + \epsilon_{ij}$, where w_i represents the true latent score, and ϵ_{ij} is the error observed by judge j . Then, w_i is linearly related to the covariates, assumed normal with mean containing the term $\mathbf{x}_i^T \boldsymbol{\beta}$. The normal latent response distribution is not appropriate when grade distributions are skewed or favor low/high scores over moderate scores. Since the distribution of the z_{ij} does not have a judge-specific mean, random cut-offs are necessary to allow the ratings of a particular subject to vary among judges.

Savitsky and Dalal (2014) note that the assumption of intrinsic agreement among raters may be inappropriate when raters have different beliefs or perspectives that may influence their scoring behavior. They assume a DP mixture model for the judge-specific latent random vectors built from a mixture kernel defined through independent normals. Dependence is therefore introduced over the latent scores of a single rater (albeit under a restrictive product-kernel for the mixture), but the data vectors arising from each rater are assumed independent. Under this model, it is therefore unclear how to extract inference for inter-rater agreement, which is a key inferential objective for applications of this type in the social sciences.

Similar to the approach in Savitsky and Dalal (2014), there is no notion in our model of an intrinsic true score for an individual. An overall score for an individual could be obtained by averaging in some fashion over the latent scores assigned by each rater. However, our main goal here is to obtain inference about relationships between the ordinal scores and the covariates, as well as for inter-rater agreement over both the covariate space and the scores. Our method offers a potentially useful perspective for modeling multirater agreement data, most notably with respect to the generality of the nonparametric mixture model which can accommodate complex dependence among raters and non-linear relationships with the random covariates.

We apply our method to a problem involving three expert graders who evaluate $n = 198$ student essays, each assigned a rating on an ordinal scale of 1 through 10 (these represent raters 2, 3, and 4 from the data given in Chapter 5 of Johnson and Albert, 1999). Average word length and total number of essay words are used as the $p = 2$ covariates, to study if they have an effect on grader ratings. The traditional measure of agreement between raters l and m in the social sciences, the polychoric correlation $\rho_{lm} = \text{corr}(Z_l, Z_m)$, can be assessed through the covariance mixing parameters $(\Sigma_1, \dots, \Sigma_N)$. As discussed in Section 2.4, the posterior predictive distribution for ρ_{lm} can be obtained by sampling at each MCMC iteration the corresponding $(\rho_{lm,1}, \dots, \rho_{lm,N})$ with probabilities (p_1, \dots, p_N) . The polychoric correlation predictive distributions for all three pairs of raters favor more heavily positive correlations (raters 1 and 3 appear to agree most strongly), but place substantial probability on negative correlations. We can study where raters l and m tend to agree or disagree by grouping the latent continuous ratings according to the strength and direction of $\text{corr}(Z_l, Z_m)$. For instance, a plot of $E(z_{il} \mid \text{data})$ and $E(z_{im} \mid \text{data})$ arranged by $E(\text{corr}(z_{il}, z_{im}) \mid \text{data})$ reveals that raters 1 and 2 strongly agree on very low ratings, but disagree when rater 2 assigns low ratings and rater 1 high ratings. It is also the case for the other pairs of raters that they strongly agree mainly at low scores.

The model provides a variety of inferences in this multivariate ordinal regression example, which can be used to assess how ratings vary across covariates, as well as how raters behave in comparison to one another. Defining a high rating as 8 or higher, and a low rating as 3 or lower, Figure 7 plots estimates for the probability of high and low rating in terms of average word

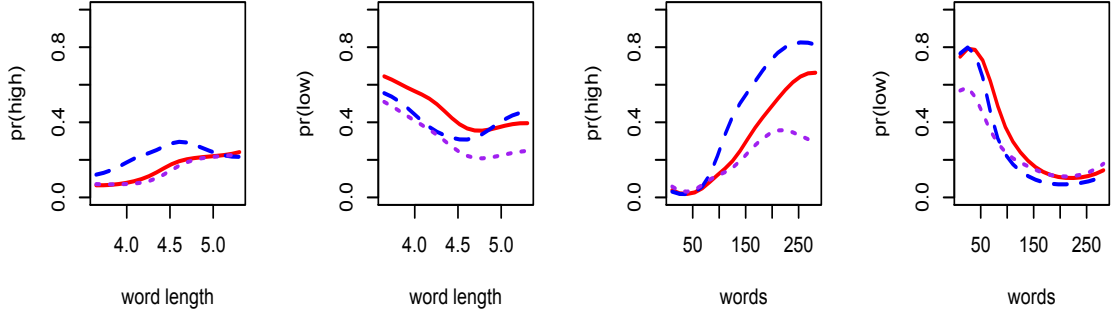


Figure 7: Multirater data. Posterior mean estimates for probability of high and low rating as a function of average word length (two left plots) and total number of words (two right plots), for raters 1 (solid red lines), 2 (dashed blue lines), and 3 (dotted purple lines).

length and total number of words. There appears to be a strong trend in rating as a function of number of words for each rater, with rater 2 in particular assigning higher ratings for essays with more words. The regression curves for high ratings associated with rater 2 are somewhat separated from raters 1 and 3, suggesting that overall rater 2 assigns higher ratings.

To identify regions of the covariate space in which raters tend to agree or disagree, Figure 8 plots estimates for the probability of perfect agreement for the three pairs of raters as a function of total number of words. This inference suggests that raters 1 and 2 agree most strongly on grades for essays with few or many words. The trend in probability of agreement is weaker for the other two pairs of raters.

Finally, to assess the strength of agreement between raters on high and low scores, we study the probability that one rater gives a high/low rating, conditional on the rating given by another rater. Table 1 includes posterior means for $\Pr(Y_l | Y_m; G)$, for $l, m \in \{1, 2, 3\}$, with Y_l and Y_m taking values of $\{8, 9, 10\}$ (high) or $\{1, 2, 3\}$ (low). Each row represents the event being conditioned on, while each column represents the event a probability is being assigned to. For example, row 1, column 3 contains the posterior mean for $\Pr(Y_2 \in \{8, 9, 10\} | Y_1 \in \{8, 9, 10\}; G)$. The cells corresponding to disagreement are highlighted with gray. The first two rows give probabilities conditional on rater 1 scores, and indicate that rater 2 has more disagreement with

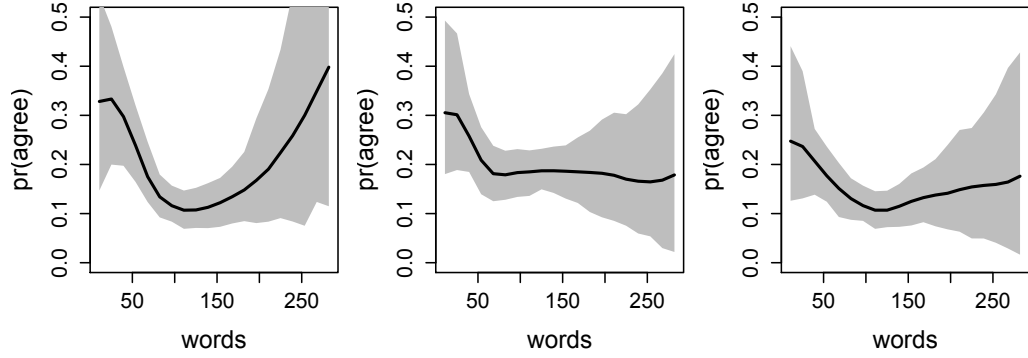


Figure 8: Multirater data. Posterior mean (solid lines) and 95% interval estimates (gray bands) for probability of agreement as a function of total number of words. The left panel corresponds to raters 1 and 2, the middle panel to raters 1 and 3, and the right panel to raters 2 and 3.

	$Y_1 = H$	$Y_1 = L$	$Y_2 = H$	$Y_2 = L$	$Y_3 = H$	$Y_3 = L$
$Y_1 = H$			0.54	0.15	0.46	0.06
$Y_1 = L$			0.13	0.48	0.04	0.51
$Y_2 = H$	0.36	0.18			0.27	0.14
$Y_2 = L$	0.08	0.56			0.07	0.41
$Y_3 = H$	0.63	0.11	0.55	0.18		
$Y_3 = L$	0.04	0.78	0.15	0.54		

Table 1: Multirater data. Posterior means for agreement and disagreement conditional probabilities for pairs of raters, with disagreement highlighted in gray (see Section 3.4 for details). H refers to high ratings of $\{8, 9, 10\}$, and L refers to low ratings of $\{1, 2, 3\}$.

rater 1 than does rater 3. The last two rows suggest that rater 2 disagrees more with rater 3 than does rater 1. Finally, from the middle two rows, we note slightly more disagreement between raters 1 and 2 than between raters 3 and 2.

4 Discussion

Seeking to expand Bayesian nonparametric methodology for ordinal regression, we have presented a fully nonparametric approach to modeling multivariate ordinal responses along with covariates. The inferential power of the framework lies in the flexible DP mixture model for the latent responses and covariates. The assumption of random covariates is appropriate for many problems, and modeling the covariates along with the latent responses accounts for dependence

or interactions among the covariates. This also allows for inference on functionals of the covariate distribution. By establishing its KL support, we have shown that the prior probability model can accommodate any mixed ordinal-continuous distribution, without imputing cut-off points or restricting the covariance matrix of the normal kernel for the DP mixture model. From a practical point of view, this is a particularly appealing feature of the modeling approach relative to the multivariate probit model and related semiparametric extensions.

The multivariate normal mixture kernel can accommodate any type of continuous covariates (using transformation as needed). Discrete ordinal covariates can also be included by introducing latent continuous variables; this was implemented in the example of Section 3.3, in which the continuous covariate income was recorded on an ordinal scale. In order to handle discrete nominal covariates, the kernel can be modified adding appropriate components to the multivariate normal density, using either a marginal or conditional specification (e.g., Taddy and Kottas, 2010).

The version of the multivariate probit model discussed in the Introduction, and the setting we consider for our model, involves a common vector of covariates $\mathbf{X} = (X_1, \dots, X_p)$ for each response vector \mathbf{Y} . That is, the covariates are not specific to particular response variables, but rather (\mathbf{Y}, \mathbf{X}) arises as a multivariate vector. An alternative version of the probit model involves p_j covariates $(X_{j,1}, \dots, X_{j,p_j})$ specific to response variable Y_j . This regression setting was described for multivariate continuous responses by Tiao and Zellner (1964), and this is the version of the multivariate binary probit model considered in Chib and Greenberg (1998).

Scenarios which make use of response specific covariates fall broadly into two categories. The first consists of problems in which only a portion of the covariate vector is thought to affect a particular response, but there may be some overlap in the subset of covariates which generate the responses. Chib and Greenberg (1998) considered a voting behavior problem of this kind in which the first of two responses was assumed to be generated by a subset of the covariates associated with the second response. This data structure can also be accommodated by modeling all covariates \mathbf{X} jointly with \mathbf{Y} , and conditioning on the relevant subset of \mathbf{X} in the regression inferences.

The other type of data structure which is occasionally handled with a multivariate regression

model with response specific covariates involves univariate ordinal responses that are related in a hierarchical/dynamic fashion. For instance, Chib and Greenberg (1998) illustrate their model with the commonly used Six Cities data, in which $\mathbf{Y} = (Y_1, \dots, Y_4)$ represents wheezing status at ages 7 through 10. Such settings are arguably more naturally approached through hierarchical/dynamic modeling. Indeed, a possible extension of the methodology developed here involves dynamic modeling for ordinal regression relationships, such that at any particular time point a unique regression relationship is estimated in a flexible fashion, while dependence is incorporated across time. We will report on this modeling extension in a future manuscript.

Appendix A: Proof of Lemma 1

Under the normal kernel, $N(\mathbf{z}, \mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, of the DP mixture model for the latent responses, $\mathbf{z} = (z_1, \dots, z_k)$, and covariates, \mathbf{x} , the kernel of the implied mixture model for the ordinal responses, $\mathbf{y} = (y_1, \dots, y_k)$, and covariates is given by $k(\mathbf{y}, \mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\gamma_{k,y_k-1}}^{\gamma_{k,y_k}} \dots \int_{\gamma_{1,y_1-1}}^{\gamma_{1,y_1}} N(\mathbf{z}, \mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) dz_1 \dots dz_k$. We assume fixed cut-off points, and $y_j \in \{1, \dots, C_j\}$ with $C_j > 2$, for $j = 1, \dots, k$.

We establish likelihood identifiability for parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in $k(\mathbf{y}, \mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. That is, from

$$k(\mathbf{y}, \mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = k(\mathbf{y}, \mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \quad (7)$$

for all (\mathbf{y}, \mathbf{x}) , with $y_j \in \{1, \dots, C_j\}$ and $\mathbf{x} \in \mathbb{R}^p$, we will obtain $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

Marginalizing over \mathbf{y} both sides of (7), we obtain $N(\mathbf{x}; \boldsymbol{\mu}_1^x, \boldsymbol{\Sigma}_1^{xx}) = N(\mathbf{x}; \boldsymbol{\mu}_2^x, \boldsymbol{\Sigma}_2^{xx})$, for all $\mathbf{x} \in \mathbb{R}^p$, and thus $\boldsymbol{\mu}_1^x = \boldsymbol{\mu}_2^x \equiv \boldsymbol{\mu}^x$, and $\boldsymbol{\Sigma}_1^{xx} = \boldsymbol{\Sigma}_2^{xx} \equiv \boldsymbol{\Sigma}^{xx}$. We also have from (7) that for each $j = 1, \dots, k$, $k(y_j \mid \mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = k(y_j \mid \mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, for all $y_j \in \{1, \dots, C_j\}$ and $\mathbf{x} \in \mathbb{R}^p$. Hence, $\Pr(Y_j \leq l \mid \mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = \Pr(Y_j \leq l \mid \mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, for $l = 1, \dots, C_j - 1$, and for all $\mathbf{x} \in \mathbb{R}^p$. That is,

$$\Phi \left(\frac{\gamma_{j,l} - \mu_1^{z_j} - \boldsymbol{\Sigma}_1^{z_j x} (\boldsymbol{\Sigma}^{xx})^{-1} (\mathbf{x} - \boldsymbol{\mu}^x)}{(\boldsymbol{\Sigma}_1^{z_j z_j} - \boldsymbol{\Sigma}_1^{z_j x} (\boldsymbol{\Sigma}^{xx})^{-1} \boldsymbol{\Sigma}_1^{x z_j})^{1/2}} \right) = \Phi \left(\frac{\gamma_{j,l} - \mu_2^{z_j} - \boldsymbol{\Sigma}_2^{z_j x} (\boldsymbol{\Sigma}^{xx})^{-1} (\mathbf{x} - \boldsymbol{\mu}^x)}{(\boldsymbol{\Sigma}_2^{z_j z_j} - \boldsymbol{\Sigma}_2^{z_j x} (\boldsymbol{\Sigma}^{xx})^{-1} \boldsymbol{\Sigma}_2^{x z_j})^{1/2}} \right), \quad (8)$$

for all $\mathbf{x} \in \mathbb{R}^p$, and for $l = 1, \dots, C_j - 1$. Because $\Phi(\cdot)$ is an increasing function, its arguments in

equation (8) must be equal. The resulting equation can be expressed in the form $\mathbf{a}^T \mathbf{x} + b = 0$.

For this equation to hold true for any \mathbf{x} , we must have $\mathbf{a} = \mathbf{0}$ and $b = 0$, which yield

$$\frac{\Sigma_1^{z_j x}}{(\Sigma_1^{z_j z_j} - \Sigma_1^{z_j x}(\Sigma^{xx})^{-1}\Sigma_1^{x z_j})^{1/2}} = \frac{\Sigma_2^{z_j x}}{(\Sigma_2^{z_j z_j} - \Sigma_2^{z_j x}(\Sigma^{xx})^{-1}\Sigma_2^{x z_j})^{1/2}}, \quad (9)$$

and

$$\frac{\gamma_{j,l} - \mu_1^{z_j} + \Sigma_1^{z_j x}(\Sigma^{xx})^{-1}\mu^x}{(\Sigma_1^{z_j z_j} - \Sigma_1^{z_j x}(\Sigma^{xx})^{-1}\Sigma_1^{x z_j})^{1/2}} = \frac{\gamma_{j,l} - \mu_2^{z_j} + \Sigma_2^{z_j x}(\Sigma^{xx})^{-1}\mu^x}{(\Sigma_2^{z_j z_j} - \Sigma_2^{z_j x}(\Sigma^{xx})^{-1}\Sigma_2^{x z_j})^{1/2}}, \quad (10)$$

for $l = 1, \dots, C_j - 1$. Using (9), (10) can be expressed as $(\gamma_{j,l} - \mu_1^{z_j})\Sigma_2^{z_j x} = (\gamma_{j,l} - \mu_2^{z_j})\Sigma_1^{z_j x}$, for each $j = 1, \dots, k$. Working with 2 of these $C_j - 1$ equations, the system can be shown to have solution $\Sigma_1^{z_j x} = \Sigma_2^{z_j x} \equiv \Sigma^{z_j x}$. Then, from (10) and (9), we obtain $\mu_1^{z_j} = \mu_2^{z_j} \equiv \mu^{z_j}$ and $\Sigma_1^{z_j z_j} = \Sigma_2^{z_j z_j} \equiv \Sigma^{z_j z_j}$, respectively.

Notice that we required 2 of the $C_j - 1$ equations of the form in (10) to arrive at this solution. If $C_j = 2$ for some j , we are unable to identify the full covariance matrix Σ . In this case, if we fix $\Sigma^{z_j z_j}$, we can identify μ^{z_j} and $\Sigma^{z_j x}$, as in DeYoreo and Kottas (2014). Although we do not require free cut-offs here due to the flexibility provided by the mixture, if $C_j > 3$, the cut-off points $\gamma_{j,3}, \dots, \gamma_{j,C_j-1}$ are also identifiable.

Finally, we need to establish identifiability for $\Sigma^{z_j z_{j'}}$, where $j \neq j'$. To this end, note that (7) implies $k(y_j, y_{j'}; \mu_1, \Sigma_1) = k(y_j, y_{j'}; \mu_2, \Sigma_2)$, for any $j, j' \in \{1, \dots, k\}$, with $j \neq j'$. Hence, for any $y_j \in \{1, \dots, C_j\}$ and $y_{j'} \in \{1, \dots, C_{j'}\}$, $\int_{\gamma_{j',y_{j'}-1}}^{\gamma_{j',y_{j'}}} \int_{\gamma_{j,y_j-1}}^{\gamma_{j,y_j}} N((z_j, z_{j'})^T; (\mu^{z_j}, \mu^{z_{j'}})^T, \mathbf{V}_1) dz_j dz_{j'} = \int_{\gamma_{j',y_{j'}-1}}^{\gamma_{j',y_{j'}}} \int_{\gamma_{j,y_j-1}}^{\gamma_{j,y_j}} N((z_j, z_{j'})^T; (\mu^{z_j}, \mu^{z_{j'}})^T, \mathbf{V}_2) dz_j dz_{j'}$. Here, matrices \mathbf{V}_1 and \mathbf{V}_2 have the same diagonal elements (given by $\Sigma^{z_j z_j}$ and $\Sigma^{z_{j'} z_{j'}}$) and off-diagonal element given by $\Sigma_1^{z_j z_{j'}}$ and $\Sigma_2^{z_j z_{j'}}$, respectively. We will therefore obtain $\Sigma_1^{z_j z_{j'}} = \Sigma_2^{z_j z_{j'}}$, for $j \neq j'$, if the bivariate normal distribution function is increasing in the correlation parameter ρ (assuming, without loss of generality, zero means and unit variances). To this end, note that $\frac{\partial}{\partial \rho} N((x, y)^T; \mathbf{0}, \mathbf{R}) = \frac{\partial^2}{\partial x \partial y} N((x, y)^T; \mathbf{0}, \mathbf{R})$, where \mathbf{R} has unit diagonal elements, and off-diagonal element ρ (e.g., Plackett, 1954). Therefore, $\frac{\partial}{\partial \rho} \int_{-\infty}^a \int_{-\infty}^b N((x, y)^T; \mathbf{0}, \mathbf{R}) dx dy = N((a, b)^T; \mathbf{0}, \mathbf{R}) > 0$, and thus the result is obtained.

Appendix B: Proof of Lemma 2

We first show that there exists at least one $f_0(\mathbf{x}, \mathbf{z}) \in \mathcal{D}$ for any $p_0(\mathbf{x}, \mathbf{y}) \in \mathcal{D}^*$, as defined in (6). This is related to the example in Canale and Dunson (2011) for modeling count data, in which a univariate continuous density $f_0(z)$ induces probability mass function $p_0(y)$. Consider $y_j \in \{1, \dots, C_j\}$, for $j = 1, \dots, k$, and define

$$f_0(\mathbf{x}, \mathbf{z}) = \sum_{y_1} \cdots \sum_{y_k} \frac{p_0(\mathbf{x}, y_1, \dots, y_k) \prod_{j=1}^k 1_{(\gamma_{j,y_j-1}^*, \gamma_{j,y_j}^*]}(z_j)}{\prod_{j=1}^k (\gamma_{j,y_j}^* - \gamma_{j,y_j-1}^*)}$$

where $\gamma_{j,l}^* = \gamma_{j,l}$ if $l \in \{1, \dots, C_j - 1\}$, $\gamma_{j,0}^* = b_j$, and $\gamma_{j,C_j}^* = d_j$, with $-\infty < b_j < \gamma_{j,1}$ and $\gamma_{j,C_j-1} < d_j < \infty$, for $j = 1, \dots, k$. Then, $f_0(\mathbf{x}, \mathbf{z})$ satisfies the relationship in (6).

Turning to the proof of the lemma, let $\text{KL}(f_0, f) = \int f_0(w) \log(f_0(w)/f(w)) dw$ be the KL divergence between densities f_0 and f . Based on the chain rule for relative entropy,

$$\text{KL}(f_0(\mathbf{x}, \mathbf{z}), f(\mathbf{x}, \mathbf{z})) = \text{KL}(f_0(\mathbf{x}), f(\mathbf{x})) + \text{KL}(f_0(\mathbf{z} | \mathbf{x}), f(\mathbf{z} | \mathbf{x}))$$

where the KL divergence between conditional densities is defined as $\text{KL}(f_0(\mathbf{z} | \mathbf{x}), f(\mathbf{z} | \mathbf{x})) = \int f_0(\mathbf{x}) \{ \int f_0(\mathbf{z} | \mathbf{x}) \log(f_0(\mathbf{z} | \mathbf{x})/f(\mathbf{z} | \mathbf{x})) d\mathbf{z} \} d\mathbf{x}$. Hence, $\text{KL}(f_0(\mathbf{x}), f(\mathbf{x})) \leq \text{KL}(f_0(\mathbf{x}, \mathbf{z}), f(\mathbf{x}, \mathbf{z}))$, and thus, for any $\epsilon > 0$, $K_\epsilon(f_0(\mathbf{x}, \mathbf{z})) \subseteq K_\epsilon(f_0(\mathbf{x})) = \{f(\mathbf{x}, \mathbf{z}) : \text{KL}(f_0(\mathbf{x}), f(\mathbf{x})) < \epsilon\}$. Using the KL property of the prior model for $f(\mathbf{x}, \mathbf{z})$, $\mathcal{P}(K_\epsilon(f_0(\mathbf{x}))) \geq \mathcal{P}(K_\epsilon(f_0(\mathbf{x}, \mathbf{z}))) > 0$, such that the prior assigns positive probability to all KL neighborhoods of the true covariate density $f_0(\mathbf{x})$.

The proof relies on the following inequality for two densities $g_1(\mathbf{t})$ and $g_2(\mathbf{t})$, where $\mathbf{t} \in \mathbb{R}^s$, and for general subsets A_1, \dots, A_s of \mathbb{R} ,

$$\int_{A_s} \cdots \int_{A_1} g_1(\mathbf{t}) \log \left(\frac{g_1(\mathbf{t})}{g_2(\mathbf{t})} \right) d\mathbf{t} \geq \int_{A_s} \cdots \int_{A_1} g_1(\mathbf{t}) d\mathbf{t} \times \log \left(\frac{\int_{A_s} \cdots \int_{A_1} g_1(\mathbf{t}) d\mathbf{t}}{\int_{A_s} \cdots \int_{A_1} g_2(\mathbf{t}) d\mathbf{t}} \right). \quad (11)$$

To prove the inequality, let $B_r = \int_{A_s} \cdots \int_{A_1} g_r(\mathbf{t}) d\mathbf{t}$, for $r = 1, 2$, such that $h_r(\mathbf{t}) = g_r(\mathbf{t})/B_r$, $r = 1, 2$, are densities on $A_1 \times \dots \times A_s$. Then, the left-hand-side of (11) can be written as $B_1 \int_{A_s} \cdots \int_{A_1} h_1(\mathbf{t}) \log \left(\frac{B_1 h_1(\mathbf{t})}{B_2 h_2(\mathbf{t})} \right) d\mathbf{t} = B_1 \log(B_1/B_2) + B_1 \int_{A_s} \cdots \int_{A_1} h_1(\mathbf{t}) \log \left(\frac{h_1(\mathbf{t})}{h_2(\mathbf{t})} \right) d\mathbf{t} \geq B_1 \log(B_1/B_2)$,

since $\int_{A_s} \cdots \int_{A_1} h_1(\mathbf{t}) \log(h_1(\mathbf{t})/h_2(\mathbf{t})) d\mathbf{t}$ is the KL divergence for densities h_1 and h_2 .

Now, consider density $f(\mathbf{x}, \mathbf{z}) \in K_{\epsilon/2}(f_0(\mathbf{x}, \mathbf{z}))$. By the chain rule, $\text{KL}(f_0(\mathbf{x}), f(\mathbf{x})) < \epsilon/2$, and $\text{KL}(f_0(\mathbf{z} | \mathbf{x}), f(\mathbf{z} | \mathbf{x})) < \epsilon/2$. We apply (11) with $s = k$, $g_1(\mathbf{t}) = f_0(\mathbf{z} | \mathbf{x})$, $g_2(\mathbf{t}) = f(\mathbf{z} | \mathbf{x})$, and $A_j = (\gamma_{j, y_{j-1}}, \gamma_{j, y_j})$, for $j = 1, \dots, k$. This yields

$$\int_{\gamma_{k, y_k-1}}^{\gamma_{k, y_k}} \cdots \int_{\gamma_{1, y_1-1}}^{\gamma_{1, y_1}} f_0(\mathbf{z} | \mathbf{x}) \log \left(\frac{f_0(\mathbf{z} | \mathbf{x})}{f(\mathbf{z} | \mathbf{x})} \right) d\mathbf{z} \geq p_0(\mathbf{y} | \mathbf{x}) \log \left(\frac{p_0(\mathbf{y} | \mathbf{x})}{p^*(\mathbf{y} | \mathbf{x})} \right) \quad (12)$$

for any configuration of values $\mathbf{y} = (y_1, \dots, y_k)$ for the ordinal responses. Here, $p^*(\mathbf{y} | \mathbf{x}) = \int_{\gamma_{k, y_k-1}}^{\gamma_{k, y_k}} \cdots \int_{\gamma_{1, y_1-1}}^{\gamma_{1, y_1}} f(\mathbf{z} | \mathbf{x}) d\mathbf{z}$, such that $p^*(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})p^*(\mathbf{y} | \mathbf{x})$. Next, we sum both sides of (12) over \mathbf{y} , then multiply both sides by $f_0(\mathbf{x})$, and finally integrate both sides over \mathbf{x} , to obtain

$$\int_{\mathbb{R}^p} f_0(\mathbf{x}) \int_{\mathbb{R}^k} f_0(\mathbf{z} | \mathbf{x}) \log \left(\frac{f_0(\mathbf{z} | \mathbf{x})}{f(\mathbf{z} | \mathbf{x})} \right) d\mathbf{z} d\mathbf{x} \geq \int_{\mathbb{R}^p} f_0(\mathbf{x}) \sum_{y_k=1}^{C_k} \cdots \sum_{y_1=1}^{C_1} p_0(\mathbf{y} | \mathbf{x}) \log \left(\frac{p_0(\mathbf{y} | \mathbf{x})}{p^*(\mathbf{y} | \mathbf{x})} \right) d\mathbf{x}$$

that is, $\text{KL}(f_0(\mathbf{z} | \mathbf{x}), f(\mathbf{z} | \mathbf{x})) \geq \text{KL}(p_0(\mathbf{y} | \mathbf{x}), p^*(\mathbf{y} | \mathbf{x}))$. Since $\text{KL}(f_0(\mathbf{z} | \mathbf{x}), f(\mathbf{z} | \mathbf{x})) < \epsilon/2$, we have $\text{KL}(p_0(\mathbf{y} | \mathbf{x}), p^*(\mathbf{y} | \mathbf{x})) < \epsilon/2$, which in conjunction with $\text{KL}(f_0(\mathbf{x}), f(\mathbf{x})) < \epsilon/2$, further implies $\text{KL}(p_0(\mathbf{x}, \mathbf{y}), p^*(\mathbf{x}, \mathbf{y})) < \epsilon$, by the chain rule. We have thus obtained

$$K_{\epsilon/2}(f_0(\mathbf{x}, \mathbf{z})) \subseteq K_{\epsilon/2}(p_0(\mathbf{y} | \mathbf{x})) \quad \text{and} \quad K_{\epsilon/2}(f_0(\mathbf{x}, \mathbf{z})) \subseteq K_{\epsilon}(p_0(\mathbf{x}, \mathbf{y}))$$

where $K_{\epsilon}(p_0(\mathbf{x}, \mathbf{y})) = \{f(\mathbf{x}, \mathbf{z}) : \text{KL}(p_0(\mathbf{x}, \mathbf{y}), p^*(\mathbf{x}, \mathbf{y})) < \epsilon\}$, from which the result emerges using the KL property of the prior model for $f(\mathbf{x}, \mathbf{z})$.

References

- Albert, J. and Chib, S. (1993), “Bayesian analysis of binary and polychotomous response data.” *Journal of the American Statistical Association*, 88, 669–679.
- Basu, S. and Mukhopadhyay, S. (2000), “Bayesian analysis of binary regression using symmetric and asymmetric links,” *The Indian Journal of Statistics Series B*, 62, 372–387.
- Boes, S. and Winkelmann, R. (2006), “Ordered response models,” *Advances in Statistical Analysis*, 90, 165–179.
- Canale, A. and Dunson, D. (2011), “Bayesian kernel mixtures for counts,” *Journal of the American Statistical Association*, 106, 1528–1539.

- Chen, M. and Dey, D. (2000), “Bayesian analysis for correlated ordinal data models,” in *Generalized Linear Models: A Bayesian Perspective*, eds. Dey, D., Ghosh, S., and Mallick, B., New York: Marcel Dekker, pp. 135–162.
- Chib, S. and Greenberg, E. (1998), “Analysis of multivariate probit models,” *Biometrika*, 85, 347–361.
- (2010), “Additive cubic spline regression with Dirichlet process mixture errors,” *Journal of Econometrics*, 156, 322–336.
- Choudhuri, N., Ghosal, S., and Roy, A. (2007), “Nonparametric binary regression using a Gaussian process prior,” *Statistical Methodology*, 4, 227–243.
- Cohen, J. (1960), “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, 20, 37–46.
- Daniels, M. and Pourahmadi, M. (2002), “Bayesian analysis of covariance matrices and dynamic models for longitudinal data,” *Biometrika*, 89, 553–566.
- DeYoreo, M. and Kottas, A. (2014), “A fully nonparametric modeling approach to binary regression,” *Bayesian Analysis*, Under revision.
- Dunson, D. and Bhattacharya, A. (2010), “Nonparametric Bayes regression and classification through mixtures of product kernels,” *Bayesian Statistics*, 9, 145–164.
- Escobar, M. and West, M. (1995), “Bayesian density estimation and inference using mixtures,” *Journal of the American Statistical Association*, 90, 577–568.
- Ferguson, T. (1973), “A Bayesian analysis of some nonparametric problems,” *The Annals of Statistics*, 1, 209–230.
- Fleiss, J. (1971), “Measuring nominal scale agreement among many raters,” *Psychological Bulletin*, 76, 378–382.
- Ghosh, J. and Ramamoorthi, R. (2003), *Bayesian Nonparametrics*, New York: Springer.
- Gill, J. and Casella, G. (2009), “Nonparametric priors for ordinal Bayesian social science models: Specification and estimation,” *Journal of the American Statistical Association*, 104, 453–464.
- Hannah, L., Blei, D., and Powell, W. (2011), “Dirichlet process mixtures of generalized linear models,” *Journal of Machine Learning Research*, 1, 1–33.
- Imai, K. and van Dyk, D. (2005), “A Bayesian analysis of the multivariate probit model using marginal data augmentation,” *Journal of Econometrics*, 124, 311–334.
- Ishwaran, H. and James, L. (2001), “Gibbs sampling methods for stick-breaking priors,” *Journal of the American Statistical Association*, 96, 161–173.
- Ishwaran, H. and Zarepour, M. (2000), “Markov chain Monte Carlo in approximate Dirichlet and Beta two-parameter process hierarchical models,” *Biometrika*, 87, 371–390.

- Johnson, V. and Albert, J. (1999), *Ordinal Data Modeling*, New York: Springer.
- Kottas, A., Müller, P., and Quintana, F. (2005), “Nonparametric Bayesian modelling for multivariate ordinal data,” *Journal of Computational and Graphical Statistics*, 14, 610–625.
- Lawrence, E., Bingham, D., Liu, C., and Nair, V. (2008), “Bayesian inference for multivariate ordinal data using parameter expansion,” *Technometrics*, 50, 182–191.
- Liu, C. (2001), “Bayesian analysis of multivariate probit models – discussion on the art of data augmentation by Van Dyk and Meng,” *Journal of Computational and Graphical Statistics*, 10, 75–81.
- Liu, J. and Wu, Y. (1999), “Parameter expansion for data augmentation,” *Journal of the American Statistical Association*, 94, 1264–1274.
- Mukhopadhyay, S. and Gelfand, A. (1997), “Dirichlet process mixed generalized linear models,” *Journal of the American Statistical Association*, 92, 633–639.
- Müller, P., Erkanli, A., and West, M. (1996), “Bayesian curve fitting using multivariate normal mixtures,” *Biometrika*, 83, 67–79.
- Newton, M., Czado, C., and Chappell, R. (1996), “Bayesian inference for semiparametric binary regression,” *Journal of the American Statistical Association*, 91, 142–153.
- Olsson, U. (1979), “Maximum likelihood estimation of the polychoric correlation coefficient,” *Psychometrika*, 44, 443–460.
- Papageorgiou, G., Richardson, S., and Best, N. (2014), “Bayesian nonparametric models for spatially indexed data of mixed type,” *arXiv:1408.1368*, stat.ME.
- Plackett, R. L. (1954), “A reduction formula for normal multivariate integrals,” *Biometrika*, 41, 351–360.
- Savitsky, T. and Dalal, S. (2014), “Bayesian non-parametric analysis of multirater ordinal data, with application to prioritizing research goals for prevention of suicide,” *Journal of the Royal Statistical Society: Series C*, 63, 539–557.
- Sethuraman, J. (1994), “A constructive definition of Dirichlet priors,” *Statistica Sinica*, 4, 639–650.
- Shahbaba, B. and Neal, R. (2009), “Nonlinear modeling using Dirichlet process mixtures,” *Journal of Machine Learning Research*, 10, 1829–1850.
- Simonoff, J. (2003), *Analyzing Categorical Data*, New York: Springer-Verlag.
- Taddy, M. and Kottas, A. (2010), “A Bayesian nonparametric approach to inference for quantile regression,” *Journal of Business and Economic Statistics*, 28, 357–369.
- Talhok, A., Doucet, A., and Murphy, K. (2012), “Efficient Bayesian inference for multivariate probit models with sparse inverse correlation matrices,” *Journal of Computational and Graphical Statistics*, 21, 739–757.

- Tanner, M. and Young, M. (1985), “Modeling ordinal scale disagreement,” *Psychological Bulletin*, 98, 408–415.
- Tiao, G. and Zellner, A. (1964), “On the Bayesian estimation of multivariate regression,” *Journal of the Royal Statistical Society, Series B*, 26, 277–285.
- Verbeek, M. (2008), *A Guide to Modern Econometrics*, John Wiley and Sons, 3rd ed.
- Walker, S. and Mallick, B. (1997), “Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing,” *Journal of the Royal Statistical Society B*, 59, 845–860.
- Webb, E. and Forster, J. (2008), “Bayesian model determination for multivariate ordinal and binary data,” *Computational Statistics and Data Analysis*, 52, 2632–2649.
- Wu, Y. and Ghosal, S. (2008), “Kullback Leibler property of kernel mixture priors in Bayesian density estimation,” *Electronic Journal of Statistics*, 2, 298–331.